International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Development of Geoinformation Technology for Monitoring Events on the Basis of Data from Unstructured Web Resource Text

Nataliya Lytvynenko, Serhii Lienkov, Olexander Lytvynenko, Oksana Banzak, Hennadii Banzak

Abstract*: *The article studies unstructured text from the informational web resources as a source of obtaining of geospatial data about current events in real time. Authors propose a scheme of construction of the geoinformation technology for monitoring events through the data of the unstructured web resource text, that will allow to collect and visualize geospatial and descriptive information about events on specific topics automatically and in real time. Key elements of the proposed methodology are geoparsing of unstructured text, geocoding of detected coordinates or addresses, and storing of results in the geodatabase. Proposed technology allows to create a web application for automatic identification and monitoring of events (objects) by parameters of their category, location and time. For the purposes of developing this system, authors proposed to use free software (except for ArcGIS Pro). This factor can be attributed to the advantages of the proposed technology.

Keywords*: *Geoinformation System, Geospatial Data, Unstructured Text, Geoparsing, Geocoding.

I. INTRODUCTION

Huge amounts of data, that circulates on the Internet nowadays, are an important component in understanding the complete picture of events that occurs in the world. Hundreds of news agencies receive current information about ongoing events in every country of the world and accumulate it on their Internet resources on the daily basis. BBC News, Reuters, The Guardian, The Washington Post, USA Today, The New York Times, Bloomberg, CNN, The Telegraph, The Daily Mail, The Wall Street Journal, The Hindustan Times, FleetUnderground, Tentaran, Connect Gujarat, CommentWise, The Nationalist, The New York Times Magazine, Fox News, and others can be mentioned as notable examples of the world-wide news agencies [1].

However, real-time monitoring of the media and news from certain categories or certain regions of the world requires more and more time because of growing amounts of information and factual repetition of the information by various sources. Development of the new geoinformation technology for automatic processing of the unstructured text data arrays will contribute to better understanding of chronology, geography and factual nature of events that occur in the world, their operational monitoring and identifying of the most important news on the Internet.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author
N. Lytvynenko*, Research Center, Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Email: n123n@ukr.net.
S. Lienkov, Research Center, Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Email: lenkov_s@ukr.net.
O. Lytvynenko, Research Center, Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Email: s63010566s@gmail.com.
O. Banzak, Head of Department Electronics and Microsystem Technology Odesa State Academy of Technical Regulation and Quality, Odesa, Ukraine. Email: banzakoksana@gmail.com.
H. Banzak, associate professor of Department Metrology and Metrological Support Odesa State Academy of Technical Regulation and Quality. Email: banzakoksana@gmail.com.

II. RESEARCH RESULTS AND ANALYSIS

The purpose of the study is to outline practical approaches to the development of the geoformation technology for monitoring events, which will use data from unstructured text of the web resources to determine location of the event, locality, and, in some cases, specific address, including street, house number or coordinates in a certain coordinate system. Indeed, a geospatial component extracted from the text — the information about the object or process location on the Earth's surface — is one of the first steps in understanding of the interest and importance of the news for a consumer. The temporal and spatial expressions are related to the events in time and space, therefore, geolocation of the event is an important premise for their correct understanding.

In the field of geoinformation systems (GIS), term "geoparcing" designates process of identification of location names in the text. This term is known in the computational linguistics as recognition and classification of the named entities (NERC) [2].

The geospatial component can be formulated as an address of the different levels of detail or geographic coordinates of various forms of representation in the text. In the GIS theory the process of conversation of the address into coordinates is called geocoding, and the process of conversation of the geographic coordinates into address is called the reverse geocoding. Geographical coordinates, as well as the addresses, after geocoding can be used to indicate a location on the cartographic basis.

In the work with geographical toponyms the task of automatic identification of toponyms (proper names denoting actual name of the natural object or object created by human) is to calculate locations of the names of the places which were found in the text in order to provide additional information about their location, for example, geographical latitude and longitude.

Establishment of correct links between toponyms and locations is difficult due to the incompleteness of databases and a large degree of ambiguity: common words should be separated from the geographical locations (geo-ambiguity), comparison between names and locations is also ambiguous. Additionally, toponyms and boundaries may change over the time, leading to incomplete and inaccurate databases.

Process of the recognition of toponyms identifies a text interval (i.e. the start and end positions of characters), which is a toponym, and then classifies it by marking the appropriate text interval as a toponym in contrast to the person's names, names of products, etc.

Since recognition of toponyms is a special case of general recognition and classification of named entities, only one class of objects is interesting for the present study - objects, which allow to determine the location. For performing toponym recognition modern systems on the first stage segment a document D_i into a tokens sequence TOKENS.

As a rule, the task of NERC consists in marking the sequence: the solution selects the most probable label for each token. For example, from a set of labels (I – LOC, B – LOC, O), where I – LOC refers to text intervals that refer to location, B – LOC is necessary for unambiguous separation of adjacent objects of the same type, O – out of the text interval, making references to a location, the fragment of unstructured text is marked as follows:

Edinburgh	is	the	vibrant	cultural	capital	of
I-LOC	0	0	0	0	0	0

Scotland	,	perhaps	its	role	is	comparable	to
I-LOC	0	0	0	0	0	0	0

the	role	New	York	plays	in	the	US	.
0	0	I-LOC	I-LOC	0	0	0	I-LOC	0

Then, when determining toponyms, the potential references (potential locations) set is searched and a correlation function is calculated that effectively selects the correctly matched candidate, discarding alternative candidates, refering to incorrect locations.

Formalized the task can be described in next way. Suppose there is a data set that includes many documents $D=\{D_1,\ldots,D_{|D|}\}$. Each document D_i contains a tokens sequence TOKENS=(TOKEN [1],…, TOKEN [|TOKENS|]. In addition, the geographical reference book G is used, that lists all reference referred to candidates $R=\{R_1,\ldots,R_{|R|}\}$. The geographic reference entrance $G(T_i)$ for the toponym T_i is a tuple containing the object type (address, settlement, mountain, proper name, etc.) and a links set $R \subset G$ for T_i. In this case, the referents are performed by the center of location, respectively, latitude and longitude. A toponym determinant is the function F_G (lat, long) that correlates locations from a document $D_i \in D$ in which toponyms are not yet defined, to the document with the same content in which toponyms are defined, that is, where a relation from a candidates set is selected for each toponym. The relations can be represented in a variety of ways, including polygons or pair of latitude and longitude coordinates of the location center. It is believed that T_i ɑ R_j, only if the toponym T_i refers to the location that represented R_j [2].

While the form of the representation of coordinates in the text is sufficiently normalized (degrees with tenths, degrees and minutes, degrees with minutes and seconds), the form of address representation in unstructured text may contain inaccuracies and uncertainties. To make it possible to use unstructured text in the GIS, the references of places mentioned in the text should be recognized automatically and compared with the geographical coordinates of these places. This process is called a geoparsing, and the software that performs this function – geoparser.

Geoparsers, free software systems, use such definitions as the area, the population, or administrative level of territory division to obtain a relevant toponyms list, variants of the geographical names in the text. Other methods are usually used after such condition is fulfilled, since it greatly helps in retrieving a relevant list of toponyms variants for further processing. The vast majority of such methods focuses on use of toponyms, repeated in the text, to resolve the ambiguities that occur in the names of places. Some of researches focus on spatial proximity (it is assumed that the toponyms in the document may constitute a "spatial cluster"), and therefore toponymic options for occurring geographical names that minimize the average distance between all possible toponyms take precedence over distant options. This hypothesis is called "spatial minimality".

There is another category of methods. These methods use the coincidence of geographical names and are based on spatial minimality. They are also guided by the names in the spatial hierarchy (country, region or region, settlement) or the same subclass names of this hierarchy.

Widely used in the definition of toponyms based on coincidences, spatial minimality cannot always help with the identification of toponyms (for example, the hypothesis of spatial minimality is not confirmed for messages on social networks). The idea is that repeated names of places should be defined as toponyms if it is necessary to minimize the area or average distance between repeated toponym's variants. It is assumed that the smaller number of such markers, the more probability of spatial minimality is preserved. However, the minimal spatial reliable approach has not been tested empirically for documents with different geographic areas or the token's number [3-5].

Spatial minimality is one of the main heuristic approaches for constructing features in methods for determining toponyms. Many other methods have been proposed that don't consider duplicate place names, but use words in the text that don't have a spatial reference to create the language spatial and thematic models, the information clusters in knowledge bases, in combination with which toponym's variants can be considered. The probability of correct result will increase with an increase in its rating. The essence of these methods is that non-geographic entities can provide important data in the process of ambiguity eliminating of toponyms.

These methods are usually used in the work with documents with thematic models, that can be generated via other sources and used to determine toponyms. There is a problem with such methods that consists in the fact of no data to evaluate them, especially if this method is used to establish additional sources of events.

Therefore, the methods for determination of toponyms based on machine learning, and using classifiers to determine them, in contrast to models using thematic clusters or classes for working, are less used.

Several studies with training classifiers for determining toponyms and another resources (because of lack of the training data) were conducted. The results of these studies indicate a minimal improvement in the performance of some data sets (not more than a few percent) compared with the basic indicators of text similarity, while at the same time, a decrease in productivity on other data sets is reported. It is known that these proposals are not implemented.

A recent study identified five geoparsing systems: GeoTxt, Geoparser, Yahoo! PlaceSpotter, CLAVIN, and TopoCluster [6].

Also the systems for recognizing geographical names and disambiguating toponyms for two data sets were tested. The study confirmed that the examined geoparsers did not provide results that could be used as the main geocoded data source.

Therefore, the systems such as GeoTxt, Edinburgh Geoparser, CLAVIN, GeoParser.io and TopoCluster can be used as a geoparser for the developing technology equally. They have their own API, which simplifies their using for their own application.

For example, the following text is inputed to the web-geoparser geoparser.io:

“Iraqi Federal Police officers hold up a captured ISIS flag in the village of Abu Saif, 6 kilometres from Mosul on February 22, 2017 in Nineveh, northern Iraq.”

In response, a following file in GeoJSON format is obtained (the file fragment):

"type": "Feature",
"properties": {
"country": "IQ",
"confidence": 1,
"name": "Qaryat Ālbū Sayf",
"admin1": "15",
"type": "populated place"
},
"id": 99471,
"geometry": {
"type": "Point",
"coordinates": [
43.16208,
36.27306
]
}

The GeoJSON file is displayed on the cartographic basis in the following form (Fig. 1):

Fig. 1.The automatic determination of coordinate values from unstructured text on the cartographic basis

Analysis of recent researches and publications dedicated to the geospatial component in unstructured text showed that the extension ArcGIS LocateXT by the ArcGIS Pro geographic information system of the American company ESRI allows to find the coordinate values in the text and visualize them on a cartographic basis [7].

The ArcGIS LocateXT extension allows to search for spatial locations in unstructured text data and generate point features representing those locations. Unstructured data is any text or document that contains location information including, but not limited to web pages, reports, emails, and social media.

Extract Locations tool has the capability to search through large blocks of text for locations and process many folders and files at once. Microsoft Office documents (Word, PowerPoint and Excel), Adobe PDF, XML, and HTML files are all compatible and there is no limit to the number of files that can be added.

For example, if you are reviewing news articles about earthquakes in Alaska and want to see each location mentioned in an article on a map. The following text is sent to the input:

“Alaska averages 100 earthquakes a day. The tectonics of the region are dominated by the interaction of the Pacific and North American plates. This interaction has accounted for three of the largest recorded earthquakes in history. The largest, measuring 9.2 on the Richter scale, occurred in the Prince William Sound (60.91°N, 147.34°W) on March 28th, 1964. The second largest Alaskan earthquake, measuring 8.7, occurred on February 4th, 1965, near the Rat Islands (51.25°N, 178.72°E). The third, measuring 8.6, occurred on March 9th, 1957, near the Andreanof Islands (51.50°N, 175.63°W).”

On the way out, once the tool has extracted the locations of the three earthquakes in the input text, the feature class appears in the Contents pane, and the resulting locations are visible in the map view (Fig. 2).

During an unstructured text processing Extract Locations tool search coordinates provided in known formats. At the same time, the mentioned tool has many settings for the most accurate search, including the possibility of fuzzy matching, the search for only specified coordinate formats, the limitation of the defined values number and more.

Fig. 2. Automatic display of address values from unstructured text using ArcGIS LocateXT

Each location recorded as a point in the output feature class with its coordinates. Regardless of the coordinate's original format, a standard format is used when the location is recorded in the file: DD — Decimal Degrees, DM — Decimal Minutes, DMS — Degrees Minutes Seconds, UTM — Universe Transverse Mercator, MGRS — Military Grid Reference System [8].

Coordinates of objects or events that were determined from the unstructured text in one of these ways, allows us to conduct geocoding. There are various methods of the forward and reverse geocoding. To a large extent, the choice of appropriate algorithm depends on the set and completeness of the source data on the basis of which geocoding is performed. Services such as Google Maps or Yandex.Maps provide a convenient API for geocoding, receiving map data and processing it. The interaction with these services is carried out in the HTTP protocol, the GET method is used to transfer the request parameters, as well as using JavaScript libraries that facilitate the use of the web -service.

An example of the request for obtaining geo-coordinates by postal address (a direct geocoding) using Static API Yandex.Maps:

http://geocode-maps.yandex.ru/1.x/?geocode= country,+city,+street&key=API-ключ. An example of the reverse geocoding request: http://geocode-maps.yandex.ru/1.x/653?geocode=lat,long&key= API-ключ.

As a response, the service returns data in XML format (by default) or JSON (if the request contains an additional parameter format = json) [9].

Thus, the structural-logical scheme for constructing geoinformation technology for monitoring events using data from unstructured text of web-resources includes following elements (Fig. 3): a web-crawler, the information web - resource, the text analysis subsystem, the geoparser, the geodatabase, a geocoder and the cartographic web - application.

With the help of the web crawler, the certain web - resources are monitored for the appearance of new articles at specified time intervals. When such material is found, the link to the new article along with an information source is records in the geodatabase and sends to the input of the text analysis subsystem (Fig. 4).

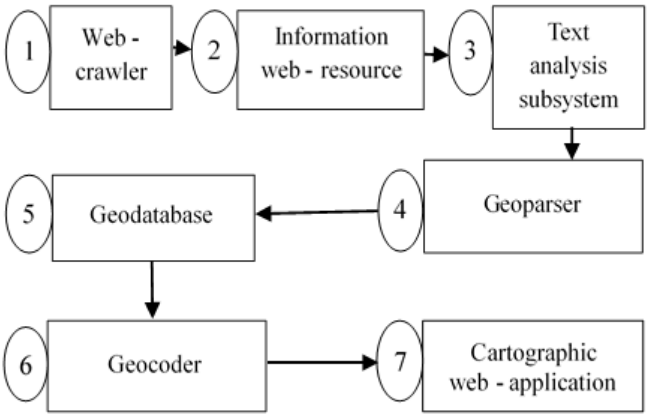

Fig. 3.The structural-logical scheme for constructing geoinformation technology for monitoring events using data from unstructured text of web – resources

Currently, there are a lot of systems, which could be used for analysis of texts and have their own API/SDK that can be used in other systems [10]. In addition, there is a significant amount of free software that allows to create and implement your own text analysis algorithms [11,12]. In this development, the text analysis subsystem should perform the following functions: the determining the name of the article and the date of publication, the reviewing the text of the article and the categorization of the article by subject (politics, economics, military, etc.). Than the text of the article is transmitted to the input of the geoparser, that finds the coordinates or addresses, converts them into coordinates and writes into the geodatabase.

For the purposes of storage of the data we propose to use PostgreSQL DBMS [13], a freely distributed object-relational database management system, one of the most developed open databases in the world and a real alternative to commercial databases. An important advantage of the proposed service is availability of additional modules that facilitate solution of geocoding problems. PostGIS [14] is a free GIS library that allows to work with geographic features and functions in a PostgreSQL database. PostGIS conforms to OpenGIS standards developed by the Open Geographic Community (OGC).

A geocoder is used to display coordinates in the form of spatially oriented points with addresses on a map (for example, Google Geocoding API or Yandex.Maps (Static API).

The results are displayed in the web-application. The following protocols and approaches in the implementation of web services are the most common: XML-RPC, SOAP, REST. As a rule, for web-services designed for searching and receiving information, the REST approach is better adapted. As a cartographic basis, a WMS service is used with a map of the required territory (for example, OSM). Using information from the geodatabase, it's possible to call additional information for the layer of objects (events), as well as apply.

III. CONCLUSION

Practical recommendations proposed in this article are aimed at implementing geoinformation technology for monitoring events using data from the unstructured text of web-resources, that will allow to collect and visualize geospatial and descriptive information about events on specific topics automatically and in real time.

For the purposes of developing this system, authors proposed to use free software (except for ArcGIS Pro). This factor can be attributed to the advantages of the proposed technology. In addition, this solution is versatile and can be used both for data of the informational resources and analysis of the data from social networks.

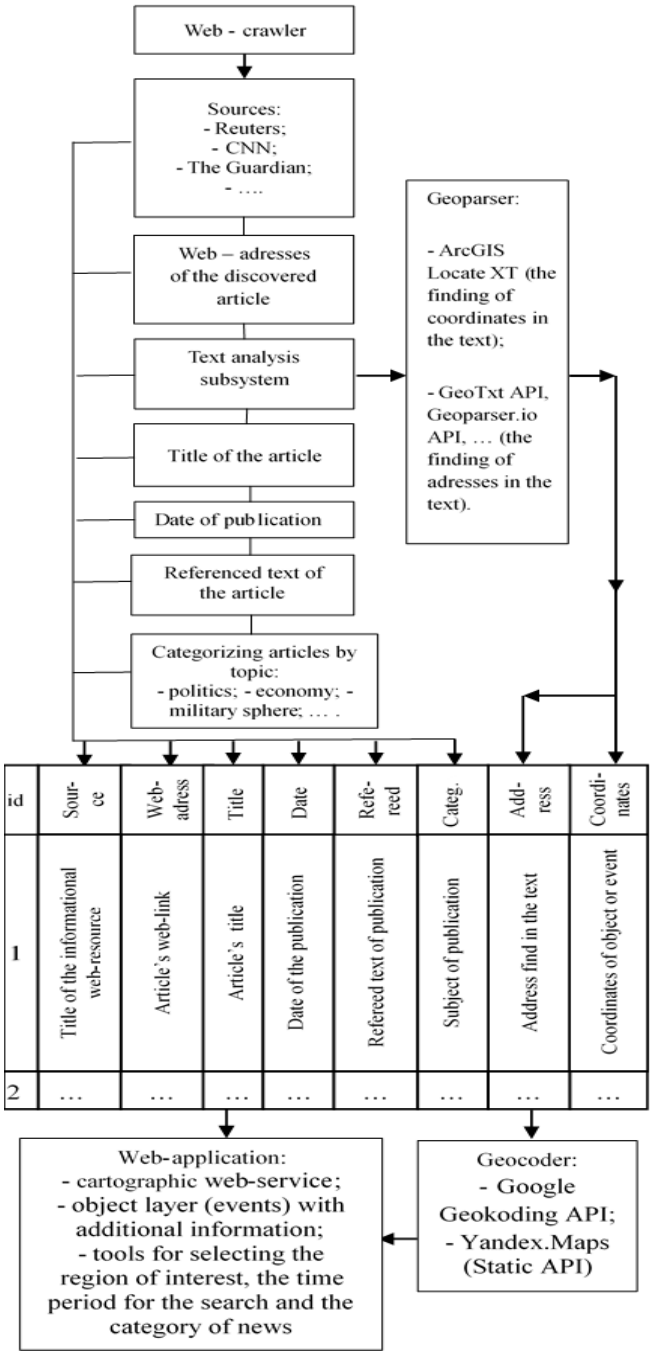

Fig. 4. The algorithm of geo-information technology work for monitoring events according to data from unstructured text of web – resources

For further researchs in this area, it is necessary to choose optimal algorithms for text analysis and geoparsing procedures, which might essentially depend on the areas of the practical application of the system.

REFERENCES

1. What are the best news websites in the world? Available at: https://en.softonic.com/solutions/what-are-the-best-news-websites-in-the-world.
2. J. L. Leidner, "Toponym Resolution in Text Annotation, Evaluation and Applications of Spatial Grounding of Place Names," Edinburgh: University of Edinburgh, 2007, 287 p.
3. D. Buscaldi, B. Magnini, "Grounding toponyms in an Italian local news corpus," Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10, Switzerland: ACM, 2010, article no. 15.
4. M. Speriosu, J. Baldridge, "Text-driven toponym resolution using indirect supervision," In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, 2013, pp. 1466–1476.
5. M. D. Lieberman, H. Same, J. Sankaranarayanan, "Geotagging with local lexicons to build indexes for textually-specified spatial data," In Proceedings of the 28th International Conference on Data Engineering, 2010, pp. 201–212.
6. A. Halterman, "Mordecai: Full text geoparsing and event geocoding," Journal of Open Source Software, no 2(9), 2017, p. 91.
7. What is LocateXT? Available: http://pro.arcgis.com/en/pro-app/help/data/locatext/extract-locations.htm#ESRI_SECTION1_1C1F4B8830AD4F78B64692E6749324F4.
8. Extract locations from Unstructured Documents. Available: http://pro.arcgis.com/en/pro-app/help/data/locatext/extract-locations-from-unstructured-documents.htm.
9. A. A. Bilchuk, D. E. Namiot, "Methods for producing geocoordinates and their application," VI International Scientific and Practical Conference "Modern Information Technologies and IT Education," Sbornik trudov. Moscow, pp. 646-658. (in Russian)
10. Top Free Software for Text Analysis, Text Mining, Text Analytics. Available: https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics.
11. The Easiest-to-Use Free/Open Source Text Analysis Software. Available: https://www.softwareadvice.com/resources/easiest-to-use-free-and-open-source-text-analysis-software.
12. PostgreSQL: The World's Most Advanced Open Source Relational Database. Available: http://www.postgresql.org.
13. What is PostGIS? Available: http://www.postgis.org.

AUTHORS PROFILE

Nataliya Lytvynenko, PhD in Technique, Senior Researcher, Research Center, Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Email: n123n@ukr.net

Serhii Lienkov, Doctor of Technical Sciences, Associate Professor, Honored Worker of Science and Technology of Ukraine, Laureate of the State Prize of Ukraine in Science and Technology, Research Center, Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Email: lenkov_s@ukr.net

Olexander Lytvynenko, PhD in Technique, Research Center, Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Email: s63010566s@gmail.com

Oksana Banzak, Doctor of Technical Sciences, Associate Professor, Head of Department Electronics and Microsystem Technology Odesa State Academy of Technical Regulation and Quality, Odesa, Ukraine. Email: banzakoksana@gmail.com

Hennadii Banzak, PhD in Technique, Associate Professor of Department Metrology and Metrological Support Odessa State Academy of Technical Regulation and Quality, Odesa, Ukraine. Email: banzakoksana@gmail.com